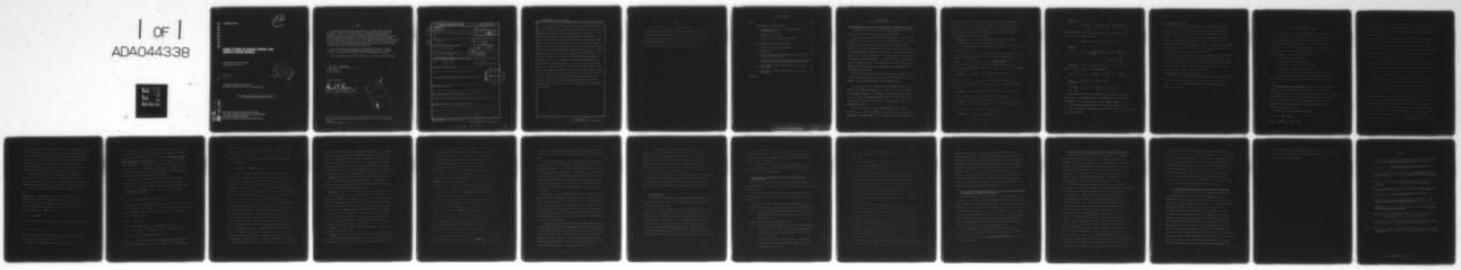
AD-A044 338 COLORADO STATE UNIV FORT COLLINS F/G 12/1
SOME STUDIES IN CODING THEORY AND SEARCH LINEAR MODELS.(U)
JUN 77 J N SRIVASTAVA, R C BOSE F33615-74-C-1198
UNCLASSIFIED AFFDL-TR-77-22 NL

1 OF 1
ADA044338



END
DATE FILMED
10-77
DDC

AFFDL-TR-77-22

(12)

SOME STUDIES IN CODING THEORY AND SEARCH LINEAR MODELS

COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523

JUNE 1977

TECHNICAL REPORT AFFDL-TR-77-22
Final Report for Period 1 June 1974 – 1 September 1976

AIR FORCE FLIGHT DYNAMICS LABORATORY
AIR FORCE WRIGHT AERONAUTICAL LABORATORIES
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Information Office (IO) and is releasable to the National Technical Information Servie (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

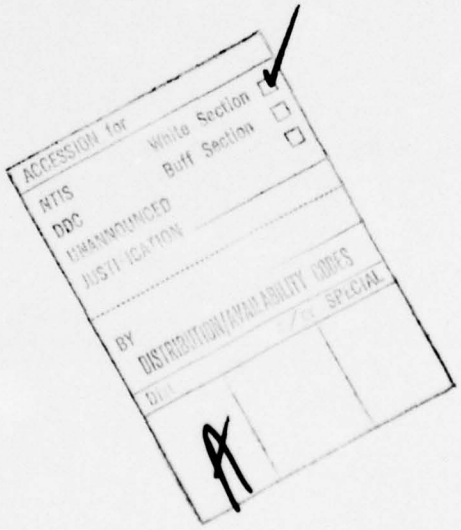
H. Leon Harter

H. LEON HARTER
Project Engineer

FOR THE COMMANDER:

Howard L. Farmer

HOWARD L. FARMER, Colonel, USAF
Chief, Structural Mechanics Division



Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

AIR FORCE/56780/29 August 1977 — 300

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFFDL-TR-77-22	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: (rept.)	
4. TITLE (and Subtitle): SOME STUDIES IN CODING THEORY AND SEARCH LINEAR MODELS.	5. TYPE OF REPORT & PERIOD COVERED: Final 1 June 1974 - 1 September 1976.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): J. N. Srivastava, R. C. Bose	8. CONTRACT OR GRANT NUMBER(s): F33615-74-1198
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Colorado State University, Fort Collins, CO 80523	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 7071-02-09, 42
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Flight Dynamics Laboratory / FBRD, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio 45433	12. REPORT DATE: June 1977
	13. NUMBER OF PAGES: 19
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same (12) 26p.	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Several related papers will be published in various journals.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear models, missing data, factor screening, optimal designs, factorial designs, search linear models, diagnosis of patients, industrial psychology

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Work was conducted on codes, designs, multivariate analysis, search linear models, and search designs. An (N + 1, k + 1) linear code of minimum weight m + 2 is a linear code in which each word is of length N + 1, the number of information symbols is k + 1, and each word has Hamming weight m + 2 or more.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

088 300

1B

If $2t + d = m + 1$, then such a code can correct up to t errors, and detect up to $t + d$ errors [1]. Here we have studied the problem of determining the function $F(m,N,k,s)$, the number of such codes. Given a design D, a subset of its treatments and blocks may form another design called a subdesign of D. Properties of some classes of designs with subdesigns have been studied. An exposition of the early history of multivariate analysis up to 1940 has been given. Consider a search linear model: $E(y) = Ap + Bq$, $V(y) = \sigma I$, where $y(N \times 1)$ is a vector of observations, $A(N \times a)$ and $B(N \times b)$ are known matrices, σ is a known or unknown non-negative constant, $p(a \times 1)$ is a vector of unknown parameters, and $q(b \times 1)$ is a vector of parameters about which partial information is available. It is known that there exists an integer k (known or unknown) such that at most k elements of q are non-zero (with their values unknown), the rest being negligible. The problem is to search the non-zero elements of q and make inferences on these and on the elements of p. In this report, we summarize the contents of a Ph.D. thesis in which studies are made on certain methods of search, and the probability of correct search under such methods. Studies on the application of search models to Factor Screening Designs, Diagnosis of Patients, and Industrial Psychology are briefly summarized, and so are certain studies on Optimal Factorial Designs and Missing Data Techniques.

PREFACE

This report constitutes the final report for the research work done under Contract No. F33615-74-1198 of the Air Force Flight Dynamics Laboratory, Air Force Systems Command, United States Air Force. The research work contained in this report was wholly supported by the above contract. The above work was technically monitored by Dr. P. R. Krishnaiah and Dr. H. L. Harter; their interest in this work is greatly appreciated.

TABLE OF CONTENTS

SECTION I
Work Related to Professor Bose

(a) Enumeration of $(N + 1, k + 1)$ Linear Codes of Minimum Weight $\geq m + 2$

A k-flat in PG (N,s) is said to be of minimum weight $\geq w$ if every point of the k-flat has at least w non-zero coordinates.

Then to each k-flat in PG (N,s) with minimum weight $\geq m + 2$ there corresponds a linear $(N + 1, k + 1)$ code with weight $m + 2$. Let $F_m(N,k,s)$ denote the number of k-flats of the above type. This function then enumerates these codes. In particular a necessary and sufficient condition for the existence of at least one $(N +1, k + 1)$ linear code of minimum weight $m + 2$ is that $F_m(N,k,s) > 0$. The enumeration of this function is therefore an important problem. A complete solution of this problem would also supply an answer to the packing problem which, besides being of importance for coding theory, has applications in factorial designs and information retrieval [1].

Work on this problem was started in collaboration with Gene Berg [2] and has been continued in collaboration with Linda Rollins and is still continuing. The results obtained so far are summarized below:

For $0 \leq i \leq N$, let X_i denote the point in PG(N,s) whose vector consists of all zeroes except for a 1 in the $(i + 1)$ - th coordinate. The fundamental simplex Δ is defined as the simplex formed by the points $X_0, X_1, \ldots, X_N$. The i-flat spanned by any $i + 1$ of the vertices $X_0, X_1, \ldots, X_N$ is called an i-cell of Δ. Then $F_m(N,k,s)$ is the number of k-flats which intersect no m-cell of Δ.

Let $H = \{X_0, X_1, \ldots, X_N\}$. An n-partition of H is a family $\mathbf{A} = \{A_1, A_2, \ldots, A_a\}$ of subsets A_i of H such that $|A_i| \geq n$, $\cup A_i = H$, and every n element subset of H is contained in a unique member of $\mathbf{A}$.

A one-partition is an ordinary partition. If A and B are n-partitions of H we say that $A < B$ if and only if $|A_i \cap B_j| \geq n \rightarrow A_i \subset B_j$. With this ordering the set of all n-partitions of H is a lattice we denote by L_n. In particular L_0 is the lattice of partitions of H ordered by refinement. The following results have been obtained:

Theorem 1. If Ψ is a k-flat in PG(N,s) which intersects no m-cell of Δ, then there exists an (m + 1)-partition $A = \{A_1, A_2, \ldots, A_a\}$ of H such that Ψ intersects Δ in every (m + 1)-cell of Δ in $A_i = (i = 1,2,\ldots,a)$ and in no other m-cell of Δ.

For $A = \{A_1, A_2, \ldots, A_a\} \in L_m$ we define $M(A)$ = the number of k flats in PG(N,s) which intersect no (m - 1)-cell of Δ but do intersect every m-cell of Δ in A_i {i = 1, 2, ... , a} and perhaps other m-cells of Δ. Similarly $N(A)$ = number of k flats in PG(N,s) which intersect no (m - 1) - cell of Δ but do intersect every m-cell in Δ in A_i {i = 1,2,...,a} and in no other m-cells of Δ.

Theorem 2. $F_m(N, k, s) = N(\phi) = \sum_{\beta \in L_m} M(\beta)\, M(\phi,\beta)$, where M is the Möbius function of L_m (in the sense of Rota's theory of generalized Möbius functions) and ϕ is the m-partition of H consisting of all $\binom{N+1}{m}$ m-subsets of H.

Theorem 3. $F_0(N, k, s) = \sum_{i=1}^{k+1} (-1)^i \binom{N+1}{i} \Phi(N - i, k - i, s)$, where $\Phi(N, m, s)$ denotes as usual the number of m-flats in PG(N,s).

Theorem 4. If we set $G_0(M, D) = F_0(M - 1, M - 1 - D, s)$, where s is a fixed prime power, then the function G_0 obeys the recurrence relation

$$G_0(M,D) = G_0(M - 1, D - 1) + (s^D - 1)\, G_0(M - 1, D).$$

Theorem 5. Let $G_0^M(x) = \sum G_0(M,D)\, x^D$; then

$$G_0^M(x) = (x - 1)\, G_0^{M-1}(x) + G_0^{M-1}(sx).$$

Theorem 6. Let

$$f(M,a) = \sum_{P(M,a)} \frac{M!}{1^{\nu_1}\, 2^{\nu_2} \cdots M^{\nu_a}} \cdot \frac{M!}{\nu_1!\, \nu_2! \cdots \nu_a!},$$

where the summation is over all partitions of M into 'a' parts; then $f(M,a)$ obeys the recurrence relation

$$f(M,a) = (M - 1)\, f(M - 1, a) + f(M - 1, a - 1).$$

Theorem 7.

$$F_1(M - 1, M - 1 - D, s) = \sum_{a=0}^{M} \left[(-1)^{M+a} (s-1)^{M-a} f(M,a) \sum_{i=0}^{a-D} (-1)^i \binom{a}{i} \Phi(a-1-i, D-1, s) \right]$$

Theorem 8. If we set $G_1(M,D) = F_1(M - 1, M - 1 - D, s)$, then the function $G_1(M,D)$ obeys the recurrence relation

$$G_1(M, D) = G_1(M - 1, D - 1) + \left[(s^D - 1) - (M - 1)(s - 1) \right] G_1(M - 1, D).$$

Theorem 9. If $G_1^M(x) = \sum G_1(M,D)\, x^D$, then

$$G_1^M(x) = \left[x - (M - 1)(s - 1) - 1 \right] G_1^{M-1}(x) + G_1^{M-1}(sx).$$

These theorems give the basic properties of the function $F_m(N, k, s)$ for the cases $m = 0$ and 1, and allow us to enumerate it for any value of the parameters N, k, s by using a very simple computer program.

A paper embodying these results is being prepared. Work on the case $m = 2$ has been started and some preliminary results obtained. It is proposed to continue this work.

(b) Subdesigns of a Design

Given any design D, whether factorial or non-factorial with a given structure, a subset of its treatments and a subset of its blocks may form another design. It is of importance to study designs which have a large number of subdesigns. A subdesign will lead to an independent estimate of error. The study of subdesigns was started under a previous Air Force contract and continued under the present contract. The paper on Baer Subdesigns of Symmetric Balanced Incomplete Designs has now appeared in "Essays in Probability and Statistics" in honor of J. Ogawa [3].

Further work on subdesigns has been continued and subdesigns of symmetric group divisible designs with the dual property have been studied, and a number of theorems obtained. In particular some applications have been made to symmetric near planes. This paper [4] has been accepted for publication in the Journal of Statistical Planning and Inference.

(c) Early History of Multivariate Analysis

A comprehensive paper [5] giving the early history of multivariate analysis, up to 1940, was prepared and formed the subject matter of the inaugural address at the 4th International Symposium on Multivariate Analysis held at Dayton, June 1975.

SECTION II
Work Related to Professor Srivastava

Most of the work under Part B was done by Professor Srivastava. For part of the work, he had one Ph.D. student, D. M. Mallenby, who at the time of this writing, is finishing up his Ph.D. dissertation. This Ph.D. dissertation was entirely supported by this Air Force contract.

The work in this section will be summarized under six different headings. These correspond to the six different topics on which significant work was performed. These topics are as follows:

1. Inference in Search Linear Models.
2. Some Studies on Missing Data Techniques.
3. Some Studies on Designs for Factor Screening.
4. Some Studies on Optimal Factorial Designs.
5. Application of Search Linear Models to the Diagnosis of Patients.
6. Application of Search Linear Models to Industrial Psychology.

In the succeeding sections 2.1-2.6, we describe the work done on these topics.

(a) Inference in Search Linear Models. In this section we shall summarize some of the main results obtained, which are contained in the Ph.D. dissertation [6] of Mallenby, Chapters 1-4. These results will be published in the form of joint papers by Srivastava and Mallenby. In order to explain the results, we shall have to recall some results from earlier papers of Srivastava [7,8]. After this, we shall briefly summarize some of the results reported in the thesis.

Consider the following general linear model:

$$(1.1a) \qquad \underline{y} = A_1^{*}\underline{\xi}_1 + A_2^{*}\underline{\xi}_2 + \underline{e},$$

$$(1.1b) \qquad \mathrm{Exp}(\underline{e}) = \underline{0}, \ \mathrm{Var}(\underline{e}) = \sigma^2 I_N,$$

where $\underline{y}(N\times 1)$ is a vector of observations, $\underline{e}(N\times 1)$ is the error vector; $A_1^*(N\times\nu_1)$, $A_2^*(N\times\nu_2)$ are known matrices, $\underline{\xi}_1(\nu_1\times 1)$ is a vector of fixed unknown parameters, and σ^2 is a known or unknown constant. About $\underline{\xi}_2(\nu_2\times 1)$, partial information is available.

The vector $\underline{\xi}_2$ consists of fixed parameters whose value is unknown. However, it is given that the elements of $\underline{\xi}_2$ are all negligible except for a set of k elements, where k is a known positive integer; however, it is not known which particular subset of k elements of $\underline{\xi}_2$ is non-negligible. In actual applications of this model, k would usually be much smaller than ν_2. This is called the search linear model with fixed effects.

With this model, the inference problem is to search out the k (possibly) non-negligible elements of $\underline{\xi}_2$, and to make inferences on these elements of $\underline{\xi}_2$ and also on the elements of $\underline{\xi}_1$. The design problem is to determine the nature of observations $\underline{y}$ (and hence the matrices A_1 and A_2), so that the search and inference problems can be handled efficiently.

Notice that in the search linear model (1.1a), there is an extra term $A_2\underline{\xi}_2$. This term is not present in ordinary general linear models, since in ordinary linear models the concept of 'search' is missing. This shows that we should expect the search model to fit real life situations better than ordinary linear models. Indeed, very often, in real life situations where one attempts to fit a linear model, one comes to a point where one feels that he has included all parameters in his model which he could lay his finger on. At the same time, he may feel that the model he has hypothesized may not fit well enough, since he knows from his experience that there must be a few more parameters which are non-negligible. However, he cannot include these parameters in his model since these parameters could be any ones out of a large set of parameters, and he does not know

exactly which parameters he should include. He may not include all of the remaining parameters because trying to estimate all of them by using the ordinary linear model, and thus determining the non-negligible parameters, would involve too many observations. Also, since he knows that most of the parameters out of the large set of parameters are negligible the ordinary linear model is not quite applicable. Clearly, the situation is described appropriately by the search linear model defined above.

The above search model was first considered by Srivastava [7]. The case when $\underline{e} = \underline{0}$ (the vector each of whose elements is zero) is called the noiseless case. This case is quite important since any difficulties arising in the noiseless case also remain present when noise is imposed.

We now present some results from Srivastava [7,8] for later use.

Theorem 1.1: Consider model (1.1a, b) under the noiseless case. A necessary and sufficient condition that the elements of $\underline{\xi}_1$ may be found exactly, and the correct non-negligible set of parameters can be searched out of $\underline{\xi}_2$ with certainty and their exact values found, is that for every (N×2k) submatrix A_{20}^{*} of $A_2^{*}(N\times\nu_2)$, we have

$$\text{Rank}(A_1^{*}:A_{20}^{*}) = \nu_1 + 2k. \tag{1.2}$$

We note that this means we must have at least $(\nu_1 + 2k)$ observations, that is,

$$N \geq (\nu_1 + 2k). \tag{1.3}$$

Examples are abundant where matrices A_1^{*} and A_2^{*} satisfy conditions (1.2), and where N attains the lower bound in (1.3).

The revelance of Theorem 1.1 to the noisy case (i.e. $\sigma^2 > 0$) is that

the rank condition (1.2) is still necessary, but it will, of course, be no longer sufficient for the inference problem to be solved. The model (1.1a,b) in which A_1^*, A_2^* satisfy (1.2) is called the strongly resolvable search linear model with fixed effects. Also, if we call T the design corresponding to the observations $\underline{y}$, then if A_1^*, A_2^* satisfy (1.2) we say T is a search design of resolving power $\{\underline{\xi}_1, \underline{\xi}_2, k\}$. Important classes of such designs have already been obtained (for example, by Srivastava and Ghosh [9]) in the context of factorial experiments.

Srivastava [8] has shown that the search and estimation problem concerning $\underline{\xi}_2$ can be separated from the estimation problem for $\underline{\xi}_1$. Accordingly, it is easier to consider the model with $\underline{\xi}_1 = \underline{0}$; this is called the pure search linear model. Set $\nu_1 = 0$, $\nu_2 = \nu$, $A_2^* = A$, and $\underline{\xi}_2 = \underline{\xi}$ in (1.1a,b) to obtain the model

$$(1.4) \qquad \underline{y} = A\underline{\xi} + \underline{e},\ \mathrm{Exp}(\underline{e}) = \underline{0},\ \mathrm{Var}(\underline{e}) = \sigma^2 I_N ,$$

where it is given that the elements of $\underline{\xi}$ are negligible except possibly for a set of at most k elements where k is a known positive integer; however, the non-negligible subset of $\underline{\xi}$ is not known. From (1.3) we have $N \geq 2k$. The rank condition becomes

$$(1.5) \qquad \mathrm{rank}(A^0) = 2k$$

for every $(N \times 2k)$ submatrix A^0 of A. There are $\binom{\nu}{k}$ distinct sets $\underline{\xi}^i$, $i = 1,2,\ldots,\binom{\nu}{k}$, of k elements of $\underline{\xi}$, and the corresponding model is

$$(1.6) \qquad \underline{y} = A_i\underline{\xi}^i + \underline{e};\ \mathrm{Exp}(\underline{e}) = \underline{0},\ \mathrm{Var}(\underline{e}) = \sigma^2 I_N,$$

where A_i is the set of k columns from A corresponding to the elements of $\underline{\xi}^i$. The search problem is to select the correct set of k non-negligible

effects, or equivalently, to pick the correct model from the set of $\binom{\nu}{k}$ models at (1.6). Notice that, in view of (1.5), each of these models is a full rank general linear model. Also, note that we must take A with columns normalized so that all the parameters are given equal weighting; in other words, we have

(1.7) $\quad A = \left[\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_\nu\right]$, where $\underline{a}_i' \underline{a}_i = 1$, $i = 1,2,\ldots,\nu$.

Four methods of search (for general k) were proposed by Srivastava [7]. Of these, two are based on the selection of that subset of k parameters (to correspond to the possibly non-zero set) which in some sense corresponds to a "small" sum of squares due to error. In the other two methods, we consider estimating the sum of squares of the non-negligible parameters. All of these methods, even under the assumption of normality, lead to rather intricate distribution theory problems. Besides the above four methods, a fifth method has been developed, where the idea is to try to dichotomize the set of parameters into two parts such that one part has the 'large' parameters, and the other does not. There is the possibility that, in general, this approach would have the merit of requiring less computation in the selection process. Below, we proceed to describe these procedures in a little more detail, and explain the results obtained.

Now we shall describe in detail the various methods of search that have been studied. We shall refer to the general search linear model with two sets of parameters $\underline{\xi}_1$, and $\underline{\xi}_2$. We shall assume that there is some set of at most k parameters, out of the set of $\underline{\xi}_2$, which is non-zero. We shall assume that the value of k is known. The methods then are as follows.

Method I. This method consists of taking a given subset of k parameters out of $\underline{\xi}_2$, and considering the ordinary linear model with $\nu_1 + k$ parameters so

obtained. Corresponding to this general linear model with $\nu_1 + k$ parameters, one could calculate the sum of the squares due to error. Corresponding to the ith set of parameters ($i = 1,\ldots,\binom{\nu}{k}$) out of $\underline{\xi}_2$, we shall get an error sum of the squares, which we denote by s_i^2. We could calculate the value of s_i^2 for each value of i. Let the integer i_0 be such that $s_{i_0}^2 = \min_i s_i^2$. Then, under Method I, the subset of k parameters in $\underline{\xi}_2$ which corresponds to the value i_0 of i will be considered as the non-negligible set of parameters.

Method II. In this method we attempt to obtain an estimate of the sum of the squares of the non-negligible parameters. That subset of k parameters out of $\underline{\xi}_2$, corresponding to which this estimate is the largest, is then considered to be the non-negligible subset.

Method III. This is a variation of Method I. In this case, we choose, arbitrarily, a number c, where c is positive. A good value of c is not known, but one advisable value might be c_0, where c_0 is a number such that there are approximately u x $\binom{\nu}{k}$ values of i such that $s_i^2 \leq c_0$. We then consider the set of all parameters in $\underline{\xi}_2$, and make a frequency distribution of these, noting the number of times any such parameters occur in those subsets of k parameters out of $\underline{\xi}_2$ for which $s_i^2 \leq c_0$. After this frequency distribution is made, we choose the k parameters of $\underline{\xi}_2$ corresponding to which the frequency is the highest. Various values of u could be used in practice. One advisable value may be u = 0.1.

Method IV. This is a variation of Method II in the same direction as Method III is for Method I.

Method V. This method is a bit too complex to be stated here in complete detail. However, the main idea is as follows. We shall discuss the idea with reference to the case when k = 1, and $\underline{\xi}_1$ is a null set, so that

$\nu_1 = 0$. The matrix A_2 will here be denoted simply by A. By considering the subspaces of the column space of A, we divide the set of parameters $\underline{\xi}$ (which here denotes $\underline{\xi}_2$ for simplicity), into two parts, $\underline{\xi}_{11}$ and $\underline{\xi}_{12}$, which are mutually exclusive and exhaustive in the sense that taken together they include all the elements of $\underline{\xi}$. Now, we construct a quadratic form $\underline{y}'Q_1\underline{y}$, and we choose a number α_1 such that if $\underline{y}'Q_1\underline{y}$ turns out to be larger than α_1, then we decide that the non-negligible parameter belongs to the subset $\underline{\xi}_{11}$, and otherwise we decide that it belongs to $\underline{\xi}_{12}$. Similarly, we construct another quadratic form $\underline{y}'Q_2\underline{y}$, and take another constant α_2, such that if $\underline{y}'Q_2\underline{y}$ turns out to be larger than α_2, then we decide that the non-negligible parameter belongs to a subset $\underline{\xi}_{21}$ of $\underline{\xi}$, and that otherwise it belongs to the complimentary subset $\underline{\xi}_{22}$. The subsets $\underline{\xi}_{21}$ and $\underline{\xi}_{22}$ are mutually exclusive and exhaustive. Now, consider the two quadratic forms taken together. Suppose that, in the particular example, the first quadratic form indicates that the non-negligible parameter belongs to $\underline{\xi}_{11}$ and the second quadratic form indicates that it belongs to $\underline{\xi}_{21}$, then we decide that the non-negligible parameter belongs to the intersection of the two subsets, namely $\underline{\xi}_{11} \cap \underline{\xi}_{21}$. Similarly, we proceed with other quadratic forms. We choose a positive integer ℓ, and construct ℓ quadratic forms. For the ith case, the set $\underline{\xi}$ is divided into two subsets, $\underline{\xi}_{i1}$ and $\underline{\xi}_{i2}$, a quadratic form $\underline{y}'Q_i\underline{y}$ is constructed and a number α_i is chosen such that if $\underline{y}'Q_1\underline{y}$ is larger than α_1 then we decide that the non-negligible parameter belongs to $\underline{\xi}_{i1}$, otherwise we decide it belongs to $\underline{\xi}_{1_2}$. Thus, for each i ($i = 1,\ldots,\ell$) we decide whether the non-negligible parameter belongs to $\underline{\xi}_{i1}$ or $\underline{\xi}_{i2}$. Finally, taking all these quadratic forms together, we decide that the non-negligible parameter belongs to $\underline{\xi}_{1j_1} \cap \ldots \cap \underline{\xi}_{\ell j_\ell}$, where $j_1,\ldots,j_\ell$ take values 1 and 2. The number ℓ is chosen in such a way that these intersections of

ℓ subsets all contain at most one parameter. Finally, decision rules are constructed to arrive at the non-negligible parameter, starting from this point.

In the first four chapters of the thesis, these methods and many small variations of these are studied. The application of these is also studied, particularly to the case of 2^n fractional factorial designs. Here, two kinds of examples are considered. One is that where $\underline{\xi}_1$ is the null set, and $\underline{\xi}_2$ contains all the parameters of the 2^n factorial. The other is the case where $\underline{\xi}_1$ corresponds to the general mean, the main effects, and the two factor interactions, and $\underline{\xi}_2$ corresponds to the remaining parameters. Such designs are called designs of resolution 5.k. Only the case k = 1 is considered.

Methods II and IV are studied very little. Some theoretical study is made for Method I, particularly in connection with its application to the factorial designs, where $\underline{\xi}_1$ is the null set. Method V is developed quite a bit, and the probability of correct search under this method is studied theoretically. How to construct the quadratic forms Q_i and the subsets $\underline{\xi}_{i1}$ and $\underline{\xi}_{i2}$ is discussed in detail. The probability of correct search is calculated theoretically for the application to the factorials, where $\nu_1 = 0$, (i.e. $\underline{\xi}$ is a null set).

A number of Monte Carlo studies have also been made regarding Methods I and V and many of their variations, particularly with reference to the two applications mentioned. The probability of correct search for Method I turns out to be very high. Method V is also not too bad. The chief advantage of Method V seems to be its potential application to other cases, and its generalizability. Method I, in general, turns out to be theoretically very complex, so far as the probability of correct

search is considered. Also, in general, the calculation of $\binom{\nu}{k}$ sums of squares due to error, particularly when ν is large, and k is even moderate, seems to be a large volume of computation. On the other hand, Method V needs just a few quantities and, therefore, it seems more attractive. However, as it stands now, the question of determination of the α's is not quite solved, although one variation of Method V where α's are done away with seems to be promising.

The most heartening result seems to be the following. The probability of correct search, as shown by Monte Carlo studies, for the case of the factorials where $\underline{\xi}_1$ is not a null set, turns out to be very high. This shows that the deisgns obtained by Srivastava and Ghosh are exceedingly good.

(b) Missing Data

In Chapter V of the above thesis, some missing data techniques have been compared using Monte Carlo work. Also, for a particular case, some theoretical studies are made.

The problem is this. Suppose that we have a trivariate normal population. This population has three univariate marginals, three bivariate marginals, and, of course, one trivariate marginal which is the whole population itself. Thus, it has $2^3 - 1 = 7$ subpopulations in it, where it is considered a subpopulation of itself. Now, suppose that samples are given not only from the trivariate marginal but also from many or all of the six subpopulations. These samples may be of different sizes. Now suppose that we want to estimate the mean vector $\underline{\mu}$ (3 x 1), the correlations

ρ_{12}, ρ_{13}, ρ_{23}, the variances σ_1^2, σ_2^2, σ_3^2, and hence also the covariances of the trivariate marginal. The Monte Carlo studies show that for small samples there are a couple of methods which seem to be generally better than all the other methods studied.

Also, the theoretical study mentioned above corresponds to the case where all elements of $\underline{\mu}$ are equal, all the correlations are equal, and also all the three variances are equal.

(c) Application of Search Linear Models to Reduced-Size Factor Screening Designs

Below we give a summary of the paper on the subject which has appeared in the proceedings [10] of the Ninth International Biometric Conference, Boston, 1976, pages 139-162.

The problem of factor screening can be stated in many ways, depending upon the assumptions. In this paper subsets of the following assumptions are made.

(C1a) Out of the total of m factors, at most d factors are effective.

(C1b) All factors have the same prior probability p of being effective.

(C2) The effective variables have much greater effect than all of the unimportant variables combined. In other words, the experimental error is small.

(C3a) There are no interactions among factors.

(C3b) There are interactions among factors. However, if an interaction involving say r ($\geq$ 2) factors is "large", then, any interaction (or "main effect") involving a subset of one or more of these r factors is also large.

(C4) The "direction" of possible effects is known. In other words, we know the "sign" of any effect. (This assumption will be often made along with (C3a).)

Most often the subset of assumptions is one assumption out of (C1a) or (C1b), the assumptions (C2) and (C4), and one of the assumptions (C3a) or (C3b).

The problem is this: how to conduct an experiment with minimal size such that under an appropriate subset of the above assumptions, we are able to search for all the effective factors.

Large factor screening experiments are common in industry and biology. One biological example is the detection of a rare attribute among members of a large population. A famous example is Wasserman type of blood test. Most situations considered so far involve a large number of factors, with interactions assumed completely absent. However, in agricultural and biological work, one is often concerned with experiments where the number of factors is relatively small, and furthermore, where interactions between two or more factors may be present. This paper, therefore, deals with both types of situations.

The purpose of this paper is manifold. We first establish the connection between the general area of factor screening and the newly started field of search linear models. This leads to certain directions of development involving certain properties of zero-one matrices. The property P_t is one of these. A matrix is said to have property P_t if every set of t distinct columns of the matrix are linearly independent. Notice that the matrix could be over the real field or over finite fields. The property P_t, over finite fields, was found to be of central importance in the theory of confounded factorial designs. Later on, it was found to be of great importance also in the theory of orthogonal fractional factorial designs, and of linear error-correctiong and error-detecting codes. In this paper, we find that we need a similar property over the real field. Noting certain

connections between the real field and the finite field GF(2), it is shown that certain matrices which have been studied earlier in the context of factorial designs and/or coding theory are useful also for factor screening experiments. The next step is to consider the process of factor screening given the observations from a particular design. (In this paper the word "design" will mean merely a set of treatment combinations from a 2^m factorial. We shall always assume we have m factors, each at two levels.) Finally, multistage procedures and reduced-size designs are considered An important feature of the area, namely, the situation where observations involve errors of variation, is not considered in this paper.

(d) Comparison of Various Optimality Criteria with Respect to Balanced Optimal Factorial Designs of the 2^m Type.

This work has not yet been written up, but it formed the basis of a lecture given by Srivastava at the International Symposium on Statistics and its Applications, held at the Indian Statistical Institute, Calcutta, during December, 1974, in honor of Professor Mahalanobis. In this paper, we consider 2^m factorial designs, with $4 \leq m \leq 8$. For each value of m, a whole range of practical values of the number of runs N is considered. In various papers of the author, either alone or with Chopra, trace optimal factorial designs have been constructed. In this paper, we also consider optimality with respect to other criteria, and present a comparison.

Some main features of this work are very briefly described by Srivastava [11], which is largely a review paper contained in a volume edited by Dr. Krishnaiah, entitled Developments in Statistics, to be published by the Academic Press.

(e) Application of Search Linear Models to the Diagnosis of Patients

Although the title has been given in terms of diagnosis of patients, the statistical problem is general and is applicable in various other situations. An invited lecture on this topic was presented in the annual meeting of the Classification Society of America at Rochester in May, 1976. However, the paper has not yet been written. I am waiting for a suitable type of data to illustrate the theoretical ideas, and the paper will be written as soon as such data becomes available. Of course, the proper acknowledgement to the present Air Force contract will be made.

The statistical problem considered is as follows. It concerns the classification of an individual in terms of 'inner conditions,' when information on that individual is given on 'external variables.' For example, in case of diagnosis of disease from outward symptoms, the inner condition would correspond to the stages of disease complicated by one or more internal factors $X_1,\ldots,X_m$. The external variables would then correspond to the measurements on external symptoms, such as blood count, pulse rate, etc. Now there may be initial data available, as a result of previous detailed study, or prolonged experience. Suppose data is available for N cases, the information for the ith case being given in terms of both the internal and external variables. Values of the internal variables could be $(X_{i1},\ldots,X_{im})$, and the information on the external variables may be given as a vector $(y_{i1},\ldots,y_{ip})$. Usually, because of insufficient data, it may be that all the different possible combinations of the X's may not be available. In other words, some inner conditions may be more common than others, and information on certain inner conditions may not be available. Similarly, in certain cases some of the y observations may not be available. The approach considered here is by considering a linear model which expresses

the expected value of each y observation in terms of the corresponding X observations. The theory of search linear models is then used to estimate the various coeffecients accurately. Finally, after a good model between the y's and X's becomes known, using the theory of discriminant functions, including Mahalanobis distance, a method is proposed for classifying a new observation vector $(y_1^*,\ldots,y_p^*)$, into one of the various X vectors. In the talk given, the X vectors were restricted to the case where each $X_{ij} = 1$ or -1, representing the presence or absence of some internal condition. This made possible the application of the present work done on the theory of 2^m factorial designs to the present problem.

(f) Application of Search Linear Models to Industrial Psychology

I intend to write a paper on this subject later on. However, this idea developed during a visit to Lackland Air Force Base in Texas. Some of the scientists there informed the author about certain models they were trying to develop to predict job difficulty. The predictor variables were eight variables denoted by $X_1,\ldots,X_8$, and functions of these. For example, X_1 indicated the number of tasks performed, X_2 the mean difficulty level of a task, from nine-level ratings, X_3 the same as X_2 except that it is from seven-level ratings, X_4 the task difficulty per unit time spent from nine-level ratings, X_5 the same as X_4 from seven-level ratings, X_6 the job difficulty at the average grade level, X_7 the time spent on selected tasks, and X_8 the range of task difficulty. The functions of these considered were simple sums of pairs of a few of these or squares of some of them or products of some of them. In all, the scientists had studied 14 functions which they denoted respectively by $Z_1,\ldots,Z_{14}$. The following question arises. Are there other important functions $\{Z_i\}$?

Could it be that there are functions $\{Z_i\}$ which would serve better than the functions that they have taken? An obvious answer to this is to use the theory of search linear models.

REFERENCES

1. Bose, R. C. (1970). Error correcting, error detecting, and error locating codes, Essays in Probability and Statistics, in honor of Prof. S. N. Roy, The Univ. of North Carolina Press, Chapter 8, 147-174.

2. Berg, G. A. (1975). An Enumeration Problem in Finite Geometries, Ph.D. dissertation, Colorado State University.

3. Bose, R. C. and Shrikhande, S. S. (1976). Baer subdesigns of symmetric balanced incomplete block designs, Essays in Probability and Statistics, in honor of Professor J. Ogawa, Shinko Tusho Co. Ltd., Tokyo, Japan, Chapter 1, 1-15.

4. Bose, R. C. (1976). Symmetric group divisible designs with the dual property, to appear in the Journal of Statistical Planning and Inference, 1.

5. Bose, R. C. (1977). Early history of multivariate analysis, Proc. Fourth Int. Symp. Mult. Analysis, edited by P. R. Krishnaiah (to appear).

6. Mallenby, D. W. (1977). Inference for the Search Linear Model, Ph.D. dissertation, Colorado State University.

7. Srivastava, J. N. (1975). Designs for searching non-negligible effects, A Survey of Statistical Design and Linear Models, edited by J. N. Srivastava, North-Holland Publishing Company, Inc., New York, 507-519.

8. Srivastava, J. N. (1976). Some further theory of search linear models, Contributions to Applied Statist., published by the Swiss-Australian Region of Biometry Society, 249-256.

9. Srivastava, J. N. (1977). Balanced 2^m factorial designs of resolution V which allow search and estimation of one extra unknown effect $4 \leq m \leq 8$, Comm. Statist. - Theor. Meth., A 6(2), 141-166.

10. Srivastava, J. N. (1976). Smaller-sized factor screening designs through the use of search linear models, Proc. Ninth Int. Biom. Conf., published by the Biometric Society, 139-162.

11. Srivastava, J. N. (1977). A review of some recent work on discrete optimal factorial designs for statisticians and for experimenters, Developments in Statistics I, edited by P. R. Krishnaiah (to appear).

☆U.S.Government Printing Office: 1977 — 757-001/27